VIHAS ADI

AI Research Engineer — Machine Learning Engineer — Generative AI Engineer

(+1) 347-665-6612 — adivihas@gmail.com — LinkedIn — Portfolio — New York, NY, USA

Summary

AI Engineer with industry and academic experience building production-grade deep learning and Generative AI systems. Specialized in transformer architectures, diffusion models, and Retrieval-Augmented Generation (RAG) pipelines using PyTorch, LangChain, and vector databases. Strong ownership across the full ML lifecycle—from large-scale data preprocessing and model training to evaluation, experiment tracking, and scalable deployment via FastAPI, Docker, MLflow, and AWS. Proven ability to translate research-driven models into high-performance, deployable AI solutions for real-world applications.

Experience

AI Engineer

Prevail Infotech INC, Sandy, Utah, USA — July 2024 – Present

- Architected and deployed an enterprise Knowledge Intelligence Platform using Retrieval-Augmented Generation (RAG), indexing 50K+ internal documents to enable natural language querying and automated knowledge retrieval across business units.
- Designed and implemented end-to-end Generative AI workflows leveraging Large Language Models (LLMs) for enterprise automation, semantic search, and contextual Q&A systems.
- Built RAG pipelines using LangChain, FAISS/Pinecone vector databases, and HuggingFace embeddings, improving contextual response accuracy by 25–35% through optimized chunking, retrieval tuning, and metadata filtering strategies.
- Developed scalable document ingestion and preprocessing pipelines for unstructured data (PDFs, reports, internal knowledge bases), enabling semantic indexing and structured summarization.
- Reduced hallucination rates by approximately 18% through retrieval re-ranking, prompt engineering refinements, and structured output formatting with validation checks.
- Deployed containerized AI services via FastAPI and Docker on AWS EC2, handling 5K+ requests/day with sub-300ms average retrieval latency and high system reliability.
- Fine-tuned transformer-based models for domain-specific NLP tasks using parameter-efficient tuning techniques and iterative prompt optimization to improve domain alignment and inference performance.
- Implemented monitoring and evaluation frameworks to measure retrieval quality, generation relevance, and system stability using automated metrics and human feedback loops.

Machine Learning Engineer

Next Cloudwave Solutions Pvt Ltd, Hyderabad, India — June 2022 – Dec 2023

- Led development of an enterprise retail demand forecasting system processing 1M+ historical transaction records to generate SKU-level weekly predictions for inventory optimization.
- Designed and deployed scalable end-to-end ML pipelines from data ingestion and preprocessing to model serving using Python, scikit-learn, and SQL.
- Built and optimized supervised learning models (classification & regression), improving predictive accuracy by 15–22% through advanced feature engineering and hyperparameter tuning.
- Engineered robust data preprocessing workflows for large structured datasets (1M+ records), improving downstream model stability and consistency.
- Productionized ML models via RESTful APIs (FastAPI), enabling real-time and batch inference integration into client systems.
- Automated model retraining and evaluation workflows, reducing experimentation time by approximately 30% and improving reproducibility.
- Implemented performance monitoring, validation protocols, and A/B testing to continuously enhance inference quality and model reliability in production.

Technical Skills

- **Programming:** Python, SQL, Java
- **Machine Learning & AI:** Machine Learning, Deep Learning, Generative AI, Large Language Models (LLMs), NLP, Computer Vision, Retrieval-Augmented Generation (RAG), Prompt Engineering, LLM Function Calling
- **Frameworks & Libraries:** PyTorch, TensorFlow, Scikit-learn, HuggingFace Transformers, LangChain, CUDA
- **Vector Databases:** FAISS, Pinecone
- **Data Engineering:** ETL design, feature engineering, batch processing, structured & unstructured data processing
- **Evaluation:** RMSE, MAE, Dice, IoU, DVH analysis, model benchmarking
- **LLM / AI Development:** OpenAI APIs, Agentic AI systems, conversational AI
- **MLOps & Deployment:** Git, Docker, FastAPI, MLflow, REST APIs, model versioning
- **Cloud & Databases:** AWS (EC2, S3), MySQL, MongoDB

Education

M.S. in Artificial Intelligence
Yeshiva University, New York, NY, USA — Jan 2024 – Dec 2025

B.Tech in Computer Science Engineering
TKR College of Engineering & Technology, Hyderabad, India — Aug 2019 – 2023

Projects

AI Research Paper Summarizer (LangChain, RAG, LLMs)

- Designed an end-to-end LLM-powered pipeline to retrieve research papers from arXiv, parse PDFs, and generate structured summaries and peer-style reviews
- Built a RAG-based summarization system using recursive text chunking and map-reduce chains for long-form documents (20–40+ pages)
- Implemented section-aware summarization (Abstract, Methods, Results, Conclusion) improving factual alignment
- Developed peer-review generation modules producing structured JSON outputs for downstream evaluation
- Containerized and deployed the RAG pipeline using Docker and FastAPI on AWS EC2 for scalable inference.

Medical Dose Diffusion Model (PyTorch, DoseDiff, MambaVision)

- Developed and benchmarked DoseDiff, CT-Mamba, and hybrid architectures for radiotherapy dose prediction using the Open-KBP dataset (50 patients) under identical experimental settings
- Demonstrated that CT-Mamba achieved the best clinical performance after extended training, reducing Dose Score from 12.38 to 1.56 and DVH error from 7.37 to 0.71 over 670 epochs
- Showed that extended training (670 epochs) had a larger impact than dataset scaling, outperforming models trained on more data but fewer epochs
- Evaluated hybrid architectures and identified instability issues that negatively impacted clinical metrics, guiding architecture refinement.

Customer Churn Prediction (Performance Improvement)

- Improved baseline customer churn prediction performance by 20–25% (MAE/RMSE) by transitioning from Linear/Ridge models to Gradient Boosting.
- Engineered behavioral features (total spend, average spend, purchase frequency), leading to an additional 10–15% reduction in prediction error.
- Applied grid search and cross-validation to optimize model hyperparameters, improving stability and generalization across validation folds.

Ride Demand Prediction System (Machine Learning, Time Series)

- Developed forecasting pipelines using temporal and spatial features across 12+ months of data
- Achieved RMSE reduction of 12–18% compared to naive seasonal baselines
- Designed the system to support operational planning and resource allocation, enabling identification of peak-demand windows and high-utilization zones
- Translated model outputs into actionable insights, helping prioritize driver allocation and capacity planning for high-demand periods